DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

IDENTIFYING TREATMENT EFFECTS UNDER DATA COMBINATION

Yanqin Fan
University of Washington

Robert Sherman
California Institute of Technology

Matthew Shum
California Institute of Technology

SOCIAL SCIENCE WORKING PAPER 1377

May 2013

Identifying Treatment Effects under Data Combination*

Yanqin Fan
U. Washington

Robert Sherman
Caltech

Matthew Shum
Caltech

May 16, 2013

Abstract

We consider the identification of counterfactual distributions and treatment effects when the outcome variables and conditioning covariates are observed in separate datasets. Under the standard selection on observables assumption, the counterfactual distributions and treatment effect parameters are no longer point identified. However, applying the classical monotone rearrangement inequality, we derive sharp bounds on the counterfactual distributions and policy parameters of interest.

Keywords: Counterfactual distributions, treatment effects, partial identification.
JEL Codes: C14, C31

*Fan: Department of Economics, University of Washington, Box 353330, Seattle, WA 98195; email: yanqin.fan@vanderbilt.edu. Sherman and Shum: Caltech, HSS, 1200 East California Blvd., Pasadena, CA 91125; email: {sherman,mshum}@hss.caltech.edu. We are grateful to Cheng Hsiao, Sergio Firpo, Marc Henry, Chuck Manski, Kevin Song, and Jeff Wooldridge for valuable comments and discussions. We thank SangMok Lee for excellent research assistance, and seminar participants at Michigan State, USC, U. Washington, the Canadian Econometrics Study Group meetings (2011, Toronto), and the Vanderbilt conference, Identification and Inference in Microeconometrics (2012) for useful comments.

1 Introduction

In this note, we consider how to identify counterfactual distributions and treatment effects when the outcome variables and the conditioning covariates are observed in separate datasets. The need to combine variables from separate datasets arises naturally in many policy applications; these include poverty analysis in which one dataset consists of program participation and the other consists of demographic attributes, or epidemiological studies in which incidence of the disease and demographic variables are observed separately.

We consider the identification of counterfactual distributions and treatment effects under the standard unconfoundedness or selection on observables assumption. It is composed of (i) the conditional independence assumption – that is, the potential outcomes are jointly independent of the treatment conditional on a set of observed covariates and (ii) the common support assumption – that is, the propensity score is strictly between 0 and 1 for all values of the conditioning covariates. When the treatment outcomes and covariates are observed in a single dataset, it is well-known that the marginal and counterfactual distributions (and hence the average treatment effects and treatment effects for the treated) are point-identified. A voluminous literature has explored many aspects of identification, inference, and computation.[1]

When outcomes and conditioning covariates are observed in separate datasets, the aforementioned point identification results break down. Using explicit representations of the marginal and counterfactual distributions via an inverse propensity-score reweighting of the data and a continuous version of the classical monotone rearrangement inequality (see Hardy, Littlewood, and Polya (1934); Cambanis, Simons, and Stout (1976)), we obtain sharp bounds on the marginal and counterfactual distributions and policy parameters of interest, including average treatment effects (ATE) and average effects of treatment on the treated (ATT).

Recent work in the treatment effects literature have made use of the result in Cambanis, Simons, and Stout (1976) and inequalities bounding the distribution functions of a sum or difference between two random variables with fixed marginals in e.g., Frank, Nelsen, and Schweizer (1987) to evaluate distributional treatment effect parameters that depend on the joint distribution of the potential outcomes (such as the probability of a positive individual treatment effect and the median of the distribution of the individual treatment effect for the treated). They include Fan and Park (2009, 2010, 2012), Firpo and Ridder (2009), Heckman, Smith, and Clements (1997), and Fan and Zhu (2009) who adopt the selection-on-observables assumption; and Fan and Wu (2010) which

[1] See, for example, Horvitz and Thompson (1952), Rosenbaum and Rubin (1983a, b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998), Dehejia and Wahba (1999), Hirano, Imbens, and Ridder (2000), Chernozhukov, Fernandez-Val, Melly (2013), Rothe (2010, 2012), Khan and Tamer (2010), and Fortin, Lemieux, and Firpo (2010), to name only a few.

considers a class of latent threshold-crossing models. Unlike the current paper, however, these works assume that outcomes and covariates are observed in the same dataset so that the *marginal* and counterfactual marginal distributions are point identified. Extending these results, this paper establishes bounds on the distributional treatment effect parameters that depend on the *joint* distribution of the potential outcomes when the marginals are partially identified.

The literature on data combination is much smaller. Manski (2000; esp. Section 5) considers bounds for a treatment effect model when the aggregate treatment outcomes and agent demographics are separately observed.[2] Cross and Manski (2002) derive sharp bounds on the "long regression" of a dependent variable Y on two sets of discrete covariates Z_1 and Z_2, when only the conditional distributions of $Y|Z_1$ and $Z_2|Z_1$ are identified from separate datasets. Ridder and Moffitt (2007; section 3.1) discuss the use of the Frechet-Hoeffding inequality in data combination contexts.[3] Hoderlein and Stoye (2009) use the Frechet-Hoeffding inequality to bound violations of the revealed-preference axioms in a repeated cross-section context. Our main contribution here is to combine insights from the treatment effects literature with the monotone rearrangement inequality to obtain identification results for counterfactual distributions and treatment effects under data combination.

The rest of this paper is organized as follows. Section 2 introduces the modelling framework, some examples, and the unconfoundedness assumption. In Section 3, we present the main identification results. Section 4 concludes. Throughout the rest of this paper, we use $F_{A|B}(\cdot|b)$ and $f_{A|B}(\cdot|b)$ to denote the distribution function and density function of the random variable A conditional on $B = b$. For a distribution function F, we use $F^{-1}(\cdot)$ to denote its quantile function.

2 The Modelling Framework and Assumptions

We now describe our treatment effects model, which follows closely the "potential outcomes" approach of Rubin (1974). We let $D \in \{0, 1\}$ denote the two states of a binary treatment[4] and let Y_D denote the corresponding outcome variable of interest for $D = 0, 1$. Y_0 and Y_1 are considered "potential outcomes"; that is, each individual agent has treatment and control outcomes Y_1 and Y_0. However, only one of these outcomes is observed. That is, his observed outcome is $Y \equiv Y_1 D + Y_0(1 - D)$. Let Z denote additional conditioning covariates (typically demographic variables) which can affect both treatment as well as potential outcomes.

[2] The ecological inference literature also considers the partial identification problem when combining aggregate and individual-level data (e.g., Glynn and Wakefield (2010)). The two-sample IV literature has considered instrumental variables models in which the outcome and the endogenous variables are observed in separate datasets (e.g., Angrist and Krueger (1992), Inoue and Solon (2010)).

[3] For a reference on Frechet-Hoeffding inequalities, see Joe (1997).

[4] As in the examples below, these treatments can be policy interventions as well as different time periods.

As a departure from the existing literature, we assume that the variables (Y, D, Z) are not observed in a single dataset. Instead, we observe two separate datasets: (i) the *outcome* dataset contains (Y, D), while (ii) the *demographics* dataset contains (Z, D). We introduce several examples below.

Example A: Long-run returns to college attendance. This data problem arises naturally in situations when the outcome of interest is a long-run outcome which is not available immediately following the treatment. For example, consider the effect of college attendance on lifetime earnings, for which there is a very large existing empirical literature. Typically, long panels, like the PSID or NLSY, are used to assess the long-run returns to college. But recent papers using the National Longitudinal Survey of Adolescent Health ("Addhealth") dataset, which is a repeated cross-section of high school students, have uncovered many rich determinants of college attendance, including parental, classroom, and even genetic factors which are not measured in other datasets (see, for example, Shanahan et. al. (2008)).

In this example Y denotes long-run earnings, observed in the PSID, while Z denotes specific determinants of college attendance, such as whether friends go to college, measures of parental attention, also genetic factors, which are only observed in Addhealth. The treatment variable $D \in \{0, 1\}$ indicates whether a student attended college, and is observed in both the PSID and Addhealth. ■

Example B: Tax payments across household types. For answering questions about tax incidence, datasets of individual tax returns are available. But tax returns contain very little demographic information on the taxpayers. For instance, one may wish to examine how tax payments vary across household types – single households, couples without children, and households with children. Tax payments and household type are observed from tax returns, but other demographic and labor market variables which are related to both tax payments and household type, such as years of education, occupational sector and hours of work, are available in labor market datasets such as the *Current Population Survey*. In this example Y denotes tax payments, D indexes the different household types, and Z are these additional demographic variables not observable from tax returns. ■

Example C: Changes in wage distribution across time. This example is drawn from DiNardo, Fortin, and Lemieux (1996). Here D is a binary indicator for two different years: $D = 0$ for the baseline year 1988, and $D = 1$ for the counterfactual year 1979. Y_D denotes wages in year D, and DiNardo, Fortin and Lemieux focus on estimating $f_{Y_0|D}(\cdot|1)$, which they interpret as the counterfactual density of wages "if individual attributes had remained at their 1979 levels and workers had been paid according to the wage schedule observed in 1988". In this example, Z would be additional covariates which affect wages. In the case when the Z variables are observed in a dataset (e.g. US Census data) separately from wages, then the results in this paper can be used to

bound the counterfactual wage distributions.[5] ■

Next, we introduce the unconfoundedness or selection on observables assumption. It is composed of two conditions. The first corresponds to the conditional independence assumption, while the second is an assumption about the support of the propensity score.[6]

(C1) Let (Y_1, Y_0, D, Z) have a joint distribution. For all $z \in \mathcal{Z}$ (the support of Z), (Y_1, Y_0) is jointly independent of D conditional on $Z = z$.

(C2) For all $z \in \mathcal{Z}$, $0 < p(z) < 1$, $0 < p_1 < 1$, where $p(z) = \Pr(D = 1|Z = z)$ and $p_d = \Pr(D = d)$ for $d = 1, 0$.

The usual approach. When (Y, D, Z) are all observed in a single dataset (so that there is no need for data combination), it is well known that under (C1) and (C2), the marginal distributions $F_{Y_1}(y)$, $F_{Y_0}(y)$ and the counterfactual distribution function $F_{Y_0|D}(y|1)$ are identified. Specifically, $F_{Y_0|D}(y|1)$ is identified through

$$F_{Y_0|D}(y|1) = \int F_{Y_0|Z,D}(y|z,1)dF_{Z|D}(z|1) = \int F_{Y_0|Z,D}(y|z,0)dF_{Z|D}(z|1) \tag{2.1}$$

in which the second equality holds under (C1). $F_{Y_1}(y)$ and $F_{Y_0}(y)$ are identified through

$$F_{Y_d}(y) = \int F_{Y_d|Z}(y|z)dF_Z(z) = \int F_{Y_d|Z,D}(y|z,d)dF_Z(z) \text{ for } d = 0, 1. \tag{2.2}$$

Thus parameters that are functionals of $F_{Y_1|Z}(\cdot|z)$, $F_{Y_0|Z}(\cdot|z)$, $F_{Y_0|D}(\cdot|1)$, including the ATE and ATT, are also identified.

However, when (Y, D) and (Z, D) are observed in separate datasets, we face a fundamental identification problem: $F_{Y_d|Z,D}(y|z,d)$ is not point identified from the sample information, so it is easy to see from (2.1) and (2.2) that $F_{Y_1|Z}(\cdot|z)$, $F_{Y_0|Z}(\cdot|z)$, and $F_{Y_0|D}(\cdot|1)$ are not point identified. To tackle this problem, we make use of the alternative expressions for $F_{Y_1}(y)$, $F_{Y_0}(y)$ and $F_{Y_0|D}(y|1)$ in terms of inverse propensity-score weighted averages below:

$$F_{Y_1}(y) = E\left[\frac{D}{p(Z)} I\{Y \leq y\}\right], \; F_{Y_0}(y) = E\left[\frac{1-D}{1-p(Z)} I\{Y \leq y\}\right], \tag{2.3}$$

$$F_{Y_0|D}(y|1) = \frac{1}{p_1} E\left[\frac{(1-D)\,p(Z)}{1-p(Z)} I\{Y \leq y\}\right]. \tag{2.4}$$

The expectations in Eqs. (2.3) and (2.4) are not point identified from the available data. We develop sharp bounds on these quantities in the next section.

[5] Fortin, Lemieux and Firpo (2010) note the formal equivalence between evaluating counterfactual distributions and evaluating treatment effects under the unconfoundedness assumption (Conditions (C1) and (C2)). See Chernozhukov, Fernandez-Val, and Melly (2013), and Rothe (2010, 2012) for related work.

[6] See e.g., Rosenbaum and Rubin (1983a, b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2000), to name only a few.

3 Identifying Treatment Effects under Data Combination

In this section, we develop sharp bounds for the marginal and counterfactual marginal distributions of the potential outcomes Y_0, Y_1 and for functionals of these distributions, including the traditional program evaluation parameters such as the ATE and ATT. We also demonstrate how sharp bounds on the marginal and counterfactual maginal distributions can be used to obtain sharp bounds on distributional treatment effects including the probability of a positive individual treatment effect and the median of the distribution of the individual treatment effect.

Our main identification results exploit a continous version of the classical monotone rearrangement inequality in Hardy, Littlewood, and Polya (1934), a special case of Theorem 2 in Cambanis, Simons, and Stout (1976).[7] For convenience, we present it in the next lemma.

Lemma 3.1 (The Cambanis-Simons-Stout inequality)**.** *Let S and T denote two random variables with known marginal distribution functions F_S and F_T. Assume S and T have finite variances. Then*

$$\int_0^1 F_S^{-1}(1-u) F_T^{-1}(u)\, du \leq E(ST) \leq \int_0^1 F_S^{-1}(u) F_T^{-1}(u)\, du.$$

Without additional information, the bounds are sharp.

It is worth pointing out that the Cambanis-Simons-Stout inequality provides sharp bounds on $E(ST)$ when the marginal distributions of S, T are known, while an application of the Cauchy-Schwartz inequality to $E(ST)$ in this case leads to bounds that are in general not sharp. Throughout the rest of this paper, we assume Assumption (I) below holds.

Assumption (I). Let $W = 1/p(Z)$ and $V = 1/[1-p(Z)]$. Assume $Var(W) < \infty$, $Var(V) < \infty$, and $Var(V/W) < \infty$. In addition, let g denote a measurable function such that $Var(g(Y_d)) < \infty$ for $d = 1, 0$.

3.1 A General Result

Our first series of results establishes sharp bounds on the mean of $g(Y_d)$:

[7] See also Chernozhukov, Fernandez-Val, and Galichon (2010) for a recent application of monotone rearrangement to constructing quantile curves without crossing.

Theorem 3.2. *(i) Let* $\mu_d(g) \equiv E(g(Y_d))$. *Then* $\mu_d^L(g) \leq \mu_d(g) \leq \mu_d^U(g)$, *for* $d = 1, 0$ *and*

$$\begin{aligned}
\mu_1^L(g) &= E\left[D\int_0^1 F_{g(Y)|D}^{-1}(1-u|D)\,F_{W|D}^{-1}(u|D)\,du\right], \\
\mu_1^U(g) &= E\left[D\int_0^1 F_{g(Y)|D}^{-1}(u|D)\,F_{W|D}^{-1}(u|D)\,du\right], \\
\mu_0^L(g) &= E\left[(1-D)\int_0^1 F_{g(Y)|D}^{-1}(1-u|D)\,F_{V|D}^{-1}(u|D)\,du\right], \\
\mu_0^U(g) &= E\left[(1-D)\int_0^1 F_{g(Y)|D}^{-1}(u|D)\,F_{V|D}^{-1}(u|D)\,du\right].
\end{aligned}$$

Without additional information, the bounds are sharp.

(ii) Let $\mu_{d|1}(g) \equiv E(g(Y_d)|D=1)$. *Then* $\mu_{1|1}(g)$ *is identified:* $\mu_{1|1}(g) = E(Dg(Y))/p_1$ *and* $\mu_{0|1}^L(g) \leq \mu_{0|1}(g) \leq \mu_{0|1}^U(g)$, *where*

$$\begin{aligned}
\mu_{0|1}^L(g) &= \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{g(Y)|D}^{-1}(1-u|D)\,F_{\frac{V}{W}|D}^{-1}(u|D)\,du\right], \\
\mu_{0|1}^U(g) &= \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{g(Y)|D}^{-1}(u|D)\,F_{\frac{V}{W}|D}^{-1}(u|D)\,du\right].
\end{aligned}$$

Without additional information, the bounds are sharp.

Proof: Consider $\mu_1(g)$. An analogue of Eq. (2.3) gives us an expression for $\mu_1(g)$ in terms of the variables (Y, D, Z), but we cannot compute this because we do not observe the joint distribution (Y, D, Z), but only the two separate distributions of (Y, D) and (D, Z). The dataset on (D, Z) allows us to identify the propensity score $p(z)$. Then, rearranging the expression, we get

$$\mu_1(g) = E\left[\frac{D}{p(Z)}g(Y)\right] = E(Dg(Y)W) = E(DE[g(Y)W|D]).$$

The rightmost quantity here contains the term $E[g(Y)W|D]$, which is the (conditional) expectation of a product of two random variables $g(Y)$ and W, which are observed in different datasets, so that the expectation cannot be computed feasibly. However, we can apply Lemma 3.1 to obtain bounds on the expectation of their product. This leads to the bounds for $\mu_1(g)$ in part (i) of Theorem 3.2. Similarly, by using the expression: $\mu_0(g) = E\left[\frac{1-D}{1-p(Z)}g(Y)\right]$, we obtain the bounds for $\mu_0(g)$ in part (i). For part (ii), noting that $V/W = p(Z)/[1-p(Z)]$, we get: $p_1\mu_{0|1}(g) = E\left[(1-D)\left(\frac{V}{W}\right)g(Y)\right]$ and the bounds in part (ii).

The bounds for $\mu_1(g)$ are sharp, in that there exist distributions of (D, Y, W) which attain these bounds. In fact, the upper bound on $\mu_1(g)$ is achieved when, conditional on D, $(g(Y), W)$ are perfectly positively dependent on each other; the lower bound is achieved when, conditional on D, $(g(Y), W)$ are perfectly negatively dependent on each other. Analogously, the upper bound on

$\mu_0(g)$ is achieved when conditional on D, $(g(Y), V)$ are perfectly positively dependent on each other and the lower bound is achieved when conditional on D, $(g(Y), V)$ are perfectly negatively dependent on each other. $\square$

We note that $\mu_1^L(g)$ and $\mu_1^U(g)$ are identified from the sample information, as $F_{g(Y)|D}(\cdot|d)$ is identified from the first dataset, $F_{W|D}(\cdot|d)$ $\left(F_{V|D}(\cdot|d)\right)$ is identified from the second dataset, and the expectation in the expressions for $\mu_1^L(g)$ and $\mu_1^U(g)$ can be identified from either dataset (or both).

3.2 Counterfactual Distributions and Treatment Effects

Let $\Delta \equiv Y_1 - Y_0$ denote the individual treatment effect. Let μ_Δ and $\mu_{\Delta|1}$ denote, respectively, the ATE and the ATT, i.e., $\mu_\Delta = E(\Delta)$ and $\mu_{\Delta|1} = E(\Delta|D=1)$. Bounds on μ_Δ and $\mu_{\Delta|1}$ follow immediately from Theorem 3.2:

$$\mu_1^L - \mu_0^U \leq \mu_\Delta \leq \mu_1^U - \mu_0^L \quad \text{and} \tag{3.1}$$
$$\frac{1}{p_1}E[DY] - \mu_{0|1}^U \leq \mu_{\Delta|1} \leq \frac{1}{p_1}E[DY] - \mu_{0|1}^L.$$

Let $g(Y_d) = I\{Y_d \leq y\}$ in Theorem 3.2. Noting that

$$F_{I_Y|D}^{-1}(u|D) = \begin{cases} 0 & \text{for } u \in [0, 1 - F_{Y|D}(y|D)) \\ 1 & \text{for } u \in [1 - F_{Y|D}(y|D), 1] \end{cases},$$

where $I_Y = I\{Y \leq y\}$, we obtain bounds for $F_{Y_1}(y)$, $F_{Y_0}(y)$ in part (i) of Theorem 3.3 below. Bounds for the counterfactual marginal distribution function $F_{Y_0|D}(y|1)$ are obtained similarly.

Theorem 3.3. *(i) For $d = 0, 1$, we have: $F_d^L(y) \leq F_{Y_d}(y) \leq F_d^U(y)$, where*

$$\begin{aligned}
F_1^L(y) &= E\left[D\int_0^{F_{Y|D}(y|D)} F_{W|D}^{-1}(u|D)\,du\right], \\
F_1^U(y) &= E\left[D\int_{1-F_{Y|D}(y|D)}^{1} F_{W|D}^{-1}(u|D)\,du\right], \\
F_0^L(y) &= E\left[(1-D)\int_0^{F_{Y|D}(y|D)} F_{V|D}^{-1}(u|D)\,du\right], \\
F_0^U(y) &= E\left[(1-D)\int_{1-F_{Y|D}(y|D)}^{1} F_{V|D}^{-1}(u|D)\,du\right].
\end{aligned}$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

(ii) $F_{Y_1|D}(y|1)$ *is identified:* $F_{Y_1|D}(y|1) = E[DI\{Y \leq y\}]/p_1$ *and* $F_{Y_0|D}(y|1)$ *is partially identified:* $F^L_{0|D}(y|1) \leq F_{Y_0|D}(y|1) \leq F^U_{0|D}(y|1)$, *where*

$$\begin{aligned} F^L_{0|D}(y|1) &= \frac{1}{p_1} E\left[(1-D)\int_0^{F_{Y|D}(y|D)} F^{-1}_{\frac{V}{W}|D}(u|D)\,du\right] \text{ and} \\ F^U_{0|D}(y|1) &= \frac{1}{p_1} E\left[(1-D)\int_{1-F_{Y|D}(y|D)}^1 F^{-1}_{\frac{V}{W}|D}(u|D)\,du\right]. \end{aligned}$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

We note that the distribution bounds in Theorem 3.3 are not only pointwise sharp but also uniformly sharp, i.e., the upper and lower bounds are distribution functions which are attainable for specific data-generating processes. To see this, consider the bounds on $F_1(\cdot)$. Both $F^L_1(\cdot)$ and $F^U_1(\cdot)$ are distribution functions. $F^L_1(\cdot)$ is the distribution function of Y_1 when conditional on D, I_Y and W are perfectly negatively dependent on each other or equivalently Y and W are perfectly positively dependent on each other; the upper bound $F^U_1(\cdot)$ is the distribution function of Y_1 when conditional on D, Y and W are perfectly negatively dependent on each other.

The uniform sharpness of the bounds in Theorem 3.3 allows us to establish sharp bounds on monotone functionals of the marginal or counterfactual marginal distribution functions. Such functionals include the quantile treatment effects (QTE) defined as

$$QTE_u = F^{-1}_{Y_1}(u) - F^{-1}_{Y_0}(u) \text{ and } QTE_{u|1} = F^{-1}_{Y_1|D}(u|1) - F^{-1}_{Y_0|D}(u|1), u \in (0,1).$$

3.3 Distributional Treatment Effects

Under the selection-on-observables assumption, when the outcomes and covariates are observed in the same dataset, Fan and Park (2009, 2010) have established bounds on the distribution of the individual treatment effect and the distribution for the treated:

$$F_\Delta(\delta) = \Pr(\Delta \leq \delta) \text{ and } F_\Delta(\delta|D=1) = \Pr(\Delta \leq \delta|D=1).$$

These are useful when one is interested in distributional treatment effects such as the probability of a positive individual treatment effect: either $\Pr(\Delta > 0)$ or $\Pr(\Delta > 0|D=1)$, and the median of Δ. Theorem 3.3 and the lemma below adapted from Frank, Nelsen, and Schweizer (1987) allow us to establish similar results to Fan and Park (2009, 2010, 2012) in our context.

Lemma 3.4. *Let* S *and* T *denote two random variables with fixed marginal distribution functions* F_S *and* F_T. *Further let* $F_{S-T}(\delta)$ *denote the distribution function of* $(S-T)$. *Then* $F^L_{S-T}(\delta) \leq$

$F_{S-T}(\delta) \leq F^U_{S-T}(\delta)$, *where*

$$\begin{aligned} F^L_{S-T}(\delta) &= \max\left(\sup_y \left[F_S(y) - F_T(y-\delta)\right], 0\right), \\ F^U_{S-T}(\delta) &= 1 + \min\left(\inf_y \left[F_S(y) - F_T(y-\delta)\right], 0\right). \end{aligned}$$

Consider, for instance, the distribution function $F_\Delta(\delta|D=1)$. From Theorem 3.3 and the conditional version of Lemma 3.4, we have:

$$\begin{aligned} F^L_\Delta(\delta|D=1) &\leq F_\Delta(\delta|D=1) \leq F^U_\Delta(\delta|D=1), \quad \text{where} \\ F^L_\Delta(\delta|D=1) &= \max\left(\sup_y \left[F_{Y_1|D}(y|1) - F^U_{0|D}(y-\delta|1)\right], 0\right), \\ F^U_\Delta(\delta|D=1) &= 1 + \min\left(\inf_y \left[F_{Y_1|D}(y|1) - F^L_{0|D}(y-\delta|1)\right], 0\right). \end{aligned}$$

Sharp bounds on the quantile function of $F_\Delta(\delta|D=1)$ follow directly from sharp bounds on $F_\Delta(\delta|D=1)$.

4 Concluding Remarks

We consider the identification of counterfactual distributions and treatment effects when the outcome variables and conditioning covariates are observed in separate datasets. Even under the selection on observables assumption, the marginal and counterfactual marginal distributions (hence the average treatment effect parameters) are no longer point identified, and we utilize the monotone rearrangement inequality to derive sharp bounds on the counterfactual distribution and policy parameters of interest. While this note focuses exclusively on identification, a companion paper (Fan, Sherman, and Shum (2012)) considers inference in these models and includes an empirical application to predicting counterfactual voting outcomes in US elections.

Extensions of the results in this note to the case that the separate datasets contain a common covariate X, i.e., one dataset contains observations on (Y, D, X) and the other contains (D, Z, X), are straightforward.

References

[1] Angrist, J. and A. Krueger (1992), "The Effect of Age at School Entry on Educational Attainment: an Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87, 328-336.

[2] Cambanis, S. , G. Simons, and W. Stout (1976), "Inequalities for $Ek(X, Y)$ When the Marginals are Fixed," *Zeitschrift fur Wahrscheinlichkeitstheorie. Verw. Gebiete* 36, 285-294.

[3] Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2010), "Quantile and Probability Curves without Crossing," *Econometrica* 78, 1093-1125.

[4] Chernozhukov, V., I. Fernandez-Val, and B. Melly (2013), "Inference on Counterfactual Distributions," *Econometrica*, forthcoming.

[5] Cross, P. J., and C. F. Manski (2002), "Regressions, Short and Long," *Econometrica*, 70(1), 357–368.

[6] DiNardo, J., N. Fortin, and T. Lemieux (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica* 64, 1001-1044.

[7] Dehejia, R. and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062.

[8] Fan, Y. and S. Park (2009), "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets," *Advances in Econometrics: Nonparametric Econometric Methods* 25.

[9] Fan, Y. and S. Park (2010), "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory* 26, 931-951.

[10] Fan, Y. and S. Park (2012), "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments," *Journal of Econometrics* 167, 330-344.

[11] Fan, Y., R. Sherman, and M. Shum (2012), "Partial Identification of Treatment Effects under Data Combination: Inference and an Application to Counterfactual Election Prediction," Working Paper.

[12] Fan, Y. and J. Wu (2010), "Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets," *Review of Economic Studies* 77, 1002-1041.

[13] Fan, Y. and D. Zhu (2009), "Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes," Working paper, Vanderbilt University.

[14] Firpo, S. and G. Ridder (2008), "Bounds on Functionals of the Distribution of Treatment Effects," Working Paper.

[15] Fortin, N., T. Lemieux, and S. Firpo (2010), "Decomposition Methods in Economics," in David Card and Orley Ashenfelter (eds.), *Handbook of Labor Economics* 4.

[16] Frank, M. and R. Nelson, and B. Schweizer (1987), "Best-Possible Bounds on the Distribution of a Sum – a Problem of Kolmogorov," *Probability Theory and Related Fields* 74, 199-211.

[17] Glynn, A., and J. Wakefield (2010), "Ecological Inference in the Social Sciences," *Statistical Methodology* 7, 307-322.

[18] Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66, 315-331.

[19] Hardy, G., J. Littlewood, and G. Polya (1934), *Inequalities*. Cambridge: Cambridge University Press.

[20] Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

[21] Heckman, J., J. Smith, and N. Clements (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies* 64, 487-535.

[22] Hirano, K., G. W. Imbens, and G. Ridder (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *NBER Technical Working Papers* 0251, National Bureau of Economic Research, Inc.

[23] Hoderlein, S. and J. Stoye (2009), "Revealed Preferences in a Heterogeneous Population, " Working paper, Boston College.

[24] Horowitz, J.L. and C.F. Manski (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica* 63, 281-302.

[25] Horvitz, D. and D. Thompson (1952), "A Generalization of Sampling Without Replacement from a Finite Universe, " *Journal of the American Statistical Association* 47, 663-685.

[26] Inoue, A. and G. Solon (2010), "Two-sample Instrumental Variables Estimators, " *Review of Economics and Statistics* 92, 557-561.

[27] Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall.

[28] Khan, S. and E. Tamer (2010), "Irregular Estimation, Support Conditions, and Inverse Weight Estimation " *Econometrica* 78, 2021-2042.

[29] Manski, C.F. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review* 80, 319-323.

[30] Manski, C.F. (2000), "Identification Problems and Decision under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics* 95, 415-442.

[31] Ridder, G. and R. Moffitt (2007), "Econometrics of Data Combination", in *The Handbook of Econometrics* 6B, chapter 75.

[32] Rosenbaum, P. R. and D. B. Rubin (1983a), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B* 45, 212-218.

[33] Rosenbaum, P. R. and D. B. Rubin (1983b), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, 41-55.

[34] Rothe, C. (2010), "Nonparametric Estimation of Distributional Policy Effects," *Journal of Econometrics* 155, 56-70.

[35] Rothe, C. (2012), "Partial Distributional Policy Effects," *Econometrica* 80, 2269-2301.

[36] Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66, 688-701.

[37] Shanahan, M., L. Erickson, S. Vaisey, and A. Smolen (2008), "Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene DRD2," *American Journal of Sociology* 114 (Suppl.), S260-S286.